# A Novel Approach to the Study of Culture and Economics: Languages Grammatical Structures. *[Work in Progress]*

Estefania Santacreu-Vasut [†]
Amir Shoham [‡]

First Version: August 2011. This Version: October 2011

**Abstract**

This paper proposes to use languages grammatical structures as markers of culture. These are straightforward to define and measure. More importantly, they are exogenous to current economic outcomes since they are stable and slow to change, improving upon existing measures. We build a data set of seven highly stable grammatical features of languages (weighted by the number of speakers) and find strong significant correlations between grammatical structures and economic, gender and institutional variables at the country level.

**JEL Code:Z1**
**Keywords: Culture, Endogeneity, Gender, Grammatical Structures, Institutions**

---

[†] Department of Economics at ESSEC Business School and THEMA, France, santacreuvasut@essec.edu
[‡]School of Business Administration College of Management Academic Studies, Rishon LeZion, Israel, and Temple University, USA amir1s@colman.ac.il
Thanks to Victor Gay and Chunbo Liu for excellent research assistance. We are grateful to the participants of the PSE Economic History seminar, the ITFA conference 2011 and to Barry Eichengreen, Ramon Ferrer i Cancho Johanna Nichols, Gerard Roland and Anne Zribi-Hertz for helpful conversations.

*A linguist who asks Why? must be a historian* Martin Haspelmath (1999, p. 205)

North (1990, 1993) seminal work argues that culture shapes the functioning of institutions. Since then, a growing body of research has corroborated the view that culture is a determinant of institutions and of economic performance.[1] Despite the growing sense that "culture matters" economists conducting empirical research rarely use culture as an explanatory variable of economic events. The main reason is that culture is too difficult to define, measure and potentially endogenous to economic outcomes. To address these challenges, this paper proposes to use the grammatical structure of languages as markers of culture. We argue that culture has an exogenous or predetermined aspect that manifests itself in the grammatical structure of language. Since languages grammatical structures are overall stable and slow to change, they are exogenous to current economic outcomes and can be used as markers of culture. Indeed, we restricted our choice of grammatical features of languages to those considered as highly stable or stable by linguists, consistent with the idea that they capture distant past influences. This is the first comprehensive analysis of the relation between grammatical structures and socio-economic ones. We build a data set of seven grammatical variables of the languages spoken for all the countries in the world. Our data set takes into account the diversity of languages and speakers within countries. Research in biology, cognitive sciences and linguistics supports the idea that languages are neither a purely biological nor an exclusively cultural phenomena but rather a bio-cultural hybrid (Evans and Levinson, 2009). We find strong and significant correlations between grammar and economic, gender and institutional variables. Overall, these suggest that our data set may be used for a variety of applications, with the advantage that our linguistic based cultural markers are easy to define, quantify, comparable across countries and exogenous to current economic outcomes. This paper contributes to two strands of the empirical literature on culture and economics. First, we improve existing measures of culture, like Hofstedes (1980,1983) cultural dimensions, House et al (2004) GLOBE study, Schwartz (1994) and the World Value Survey. In particular, the main advantage of our approach is the exogeneity of linguistic based variables compared to survey based

[1] Among others, Licht, Goldschmidt, and Schwartz, 2007; Tabellini, 2008; Gorodnichenko and Roland, 2010; Guiso, Sapienza, and Zingales, 2010.

ones. In addition, because of the linguistic diversity within countries we can measure dominant versus minority and colonial inherited cultures. Finally, compared to existing culture databases covering a small sample of countries, biased towards western and more developed countries, ours covers the largest sample of countries, reaching 204 countries. Second, we provide a comprehensive analysis between the relation of linguistic and economic variables. While the idea that cultural knowledge is transmitted via language was raised by North (1993) in his Nobel prize lecture, its use has been sporadic and has lacked a comprehensive approach.[2] Moreover, this is the first paper to argue that the grammatical structure of language itself captures culture and correlates with economic outcomes. Among the predecessors in using linguistic variables, Licht, Goldschmidt and Schwartz (2004) use the grammar of pronouns as an instrumental variable to study how countries more tilted in favor of autonomy, egalitarianism, and mastery exhibit higher rule of law, less corruption, and more democratic accountability. They argue that languages which require the explicit use of I or you signal that a person is highlighted and autonomy is valued. Licht, Goldschmidt Schwartz (2007) also use a linguistic variable on pronoun drop as an instrument for cultural emphases on autonomy versus embeddedness. They point out a significant influence of culture (linguistic) on governance. Tabellini (2008) seek to capture a distinction between values consistent with generalized versus limited morality. Altogether, norms of generalized morality induce well-functioning institutions. To control for the possibility of reverse causality, and identify the causal impact of these values on institutional outcomes Tabellini uses the grammar of pronouns as an instrumental variable. Tang and Kevoes (2008) argue that changes in economic conditions are the source of cultural dynamics, while institutions, among them language, religion and legal origin provide the foundation for cultural stability. Guiso, Sapienza and Zingales (2009) document that trust is affected by geographical distance between two countries, their proximity, and the commonality between the two languages. Finally, Falck at al.(2010) and Cavalli-Sforza (2000) argue that language acts as a type of memory that stores information like the genome

[2]North argues that "the kind of learning that the individuals in a society acquired through time. Time in this context entails not only current experiences and learning but also the cumulative experience of past generations that is embodied in culture. Collective learning - a term used by Hayek - consists of those experiences that have passed the slow test of time and are embodied in our language, institutions, technology, and ways of doing things. It is "the transmission in time of our accumulated stock of knowledge" (Hayek 1960: 27).

does.[3] We proceed as follows. First, we discuss why using grammatical structures as markers of culture. Second we describe our data set. Third, we present correlation matrices between grammatical structures and economic, gender and institutional variables. We conclude the paper by discussing the implications of our findings.

# 1 Why Grammatical Structures?

The main advantage of using language grammatical structures results from the fact that they are inherited from the distant past and, therefore, exogenous to current economic conditions. A recent endeavor of linguistics research studies the stability of grammatical structures. In particular, Wichmann and Holman (2009) have constructed a measure of stability to analyze the linguistic features described in the World Atlas of Linguistic Structures, (Dryer and Haspelmath, 2011). They define stability as "the probability that a given language remains unchanged with respect to the feature during 1000 years, that is, the feature undergoes neither internal change nor diffusion during the interval". We preselected eleven features from this source, and further restricted our sample to the seven that they categorize as very stable (six) and stable (one). We discard the other four. For example, for expressions of time, we keep the past tense, which is a very stable feature while discard the future tense, which is unstable.[4] Our seven grammatical features are: the number of genders, the past tense, the optative, the order of subject, object, and verb, the order of adjective and noun, and exceeding comparative constructions. From an economic point of view, the stability of grammatical features is not surprising and can be related to how network externalities affect technology adoption (Shapiro and Katz ,1986). Language is a technology characterized by networks externalities, since the value of mastering a language increases in the number of speakers. Linguistic evolution can be seen as a sort of technological adoption. As Shapiro and Katz show "in the absence of sponsors, the technology superior today

[3]To back up there claim they even quote Darwins Origin of Species "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the languages now spoken around the world; and if all extinct languages, and all intermediate and slowly changing dialect, were to be included, such an arrangement would be the only possible one.

[4]They compare their finding with categorical statements in the literature regarding the stability of the features they analyze and there is a high degree of concordance.

has a strategic advantage and is likely to dominate the market". This is likely to be the case of languages, that are characterized by an absence of owner or sponsor (which they define as "an entity that has property rights to the technology and hence is willing to make investments to promote it.") While during most of the past century, linguistics, in particular the so called generativists, viewed language as part of human biology and argued that all languages are fundamentally equal in structure with only minor local differences (Chomsky 1980 and 1991), Sampson et al. (2009) points at a new wave of linguists that see languages as "institutions developed as part of a society's cultural heritage and hence as differing and evolving in their levels of complexity, just as other cultural institutions do." The study of the evolution of kinship specific pronouns in Australia by Evans (2003b) is an example of studies of how cultural forces may shape linguistic change. As Evans and Levinson point out "language diversification and hybridization works just like the evolution of biological species it is a historical process, following the laws of population biology" and that "linguistic diversity is structured very largely in phylo-genetic (cultural-historical) and geographical patterns." Christiansen and Kirby (2003) review the research on the origins and evolution of human language and argue that while one line of theory argues that grammatical structure is the product of biological adaptation, others argue that it emerges through cultural transmission of language across hundreds (or perhaps thousands) of generations of learners. Overall, current theories of coevolution point at the interplay between genetic and cultural forces to account for language diversity and change. Interestingly, there seems to be a link between genetic families and linguistic families. Recent work by Dediu (2011) suggests that, in addition to the view that languages and genes may coevolve due to similar population-level processes, leading to a correlation between the two, there may be a causal link from the later to the former.[5] Indeed, Cavalli-Sforza have used genetic distance to measure cultural distance. Yet, while genetic based measures are less arbitrary than survey measures, they are hard to interpret. They may be useful to assess cultural distance, but in which aspects this distance materializes and where does it come from remains unclear. Further, our approach improves existing survey based mea-

[5]He argues that mathematical and computational models suggest that small genetic biases of language use at the individual level can lead to language wide trajectory changes.

sures of culture as well as those based on the reconstruction of evidence on cultural practices. For example, Schwartz (1994,2004) cultural variables are based on surveys made in 1988-1998 to urban teachers to capture national culture. These may not be equally representative of countries cultures since not all societies value education equally or have generalized access to it. Hofstede (1980, 1983) measures culture by comparing employees values at the same company operating in different countries. The advantage of this strategy is that it directly controls for education, occupation and other characteristics. The caveat is that those employees representativeness of their respective country national culture may vary across countries. Other survey based cultural variables (see Globe study (House at al, 2004), World Value Survey) face similar limitations inherent to the survey based approach: endogeneity and lack of representativeness. Following an ethnographic approach, the Standard Cross-Cultural Sample, (Murdock and White 1969) contains discrete indicators of cultural practices, yet they allow to track only some form of culture evolution. From the methodological standpoint, languages grammatical structures are defined in a precise manner and have been systematically and consistently categorized across countries by linguistic researchers. This allows us to measure culture in a clear and consistent manner for a sample of 204 countries, by far the largest and most representative dataset on markers of culture available.

# 2 The Data

## 2.1 Linguistic Data: Sources and Selection Criteria

The data for the *Languages spoken and percentage of speakers* were found on three sources: the CIA world factbook, the website Ethnologue Languages of the World and a French website, L'aménagement linguistique du monde. The source for each figure is given in the data set. The data for *Diversity, number of living languages, average number of speakers per language and mean number of speakers per language* were taken from Ethnologue Languages of the world. The data for the *Grammatical structures of languages* were taken from The World Atlas of Languages Structure

Online (WALS, Dryer and Haspelmath, 2011).[6] While the WALS dataset contains many other linguistic variables, we preselected eleven based on their potential relation with cultural values and prior to knowledge of their relation with economic variables. Our approach is based on forefront research putting forward the idea that languages are the result of both genetic and cultural forces. As Evans and Levinsion (2009) argue "language is a bio-cultural hybrid, a product of intensive gene culture coevolution over perhaps the last 200,000 to 400,000 years". To capture those exogenous past cultural forces, we further restricted our sample to seven that linguists classify as highly stable or stable.

## 2.2 Our grammatical structure variables

We build a data set composed of seven grammatical structure variables based on language structural features from the WALS database. To take into account linguistic diversity within countries, we take all the languages spoken in a country and compute a weighted average of the features of each of these languages, where the weights for a given language correspond to the percentage of speakers in the country. We next describe our variables (their precise coding is available in the appendix) and how we computed weighted variables per country to take into account linguistic diversity within countries.

### 2.2.1 Description of variables

1. Number of Genders As marker of gender we use a categorical variable from zero to five depending on whether the language has none, two, three, four or five and more number of genders.[7]

2. Past tense As marker of cultural stances regarding the past, we use information on the degree

[6] There is a consensus among linguists that, despite potential limitations of the dataset, this is the best available data describing the linguistic features of languages across the globe. The names of languages in the data set are expressed as on WALS. For instance, while the dominant language for the Philippines is usually called *Filipino*, it is called *Tagalog* in the data set as on the WALS website.

[7] " a language has a gender system only if we find different agreements ultimately dependent on nouns of different types. In other words, there must be evidence for gender outside the nouns themselves" (WALS chapter 30. Although in general it is, it may not necessarily be sex based.

of existence of past specific constructions and the degree of remoteness available.[8] In particular we build a categorical variable equal to 0, 1, 2, 3 depending on whether the language has either (a) No grammatical marking of past/non-past distinction, or (b) Past/non-past distinction marked; no remoteness distinction or (c) Past/non-past distinction marked; 2-3 degrees of remoteness distinguished, or (d) Past/non-past distinction marked; at least 4 degrees of remoteness distinguished respectively.

3. The Optative As marker of self expression, we use information on whether an inflected verb form, dedicated to the expression of the wish of the speaker, exists.[9] In particular, we build a dummy variable equal to 0 or 1 depending on whether the language has either inflectional optative absent or present respectively.

4. Order of Object and Verb As marker of the value of effort versus status (or prior wealth) to acquire wealth, we use information on the order of object and verb used by default in a sentence. In particular, we build a categorical variable equal to 0.5, 0, or 1 depending on whether the language has (a) Both orders with neither order dominant, or (b) Object precedes verb (OV) or (c) Object follows verb (VO) respectively.

5. Order of Adjective and Noun As marker of creativity and innovation ,we use information on the order of adjective and noun used by default in a sentence. In particular, we build a dummy variable equal to 0.5, 0 or 1 depending on whether the language has (a) Both orders of noun and modifying adjective occur, with neither dominant, or (b) Modifying adjective precedes noun (AdjN) or (c) Modifying adjective follows noun (NAdj) respectively. [10]

6. Reason Clauses As marker of the expression of relations between facts or events we use information on the form of the verb in reason clauses. These are "defined in functional,

[8] For example, one of the richest languages in terms of past marking is Yagua, spoken in Peru. According to Payne and Payne 1990, it has five degrees of remoteness in the past. These are captured in five different verb suffixes to capture whether the action took place "a few ours previous to the time of utterance", "one day previous to the time of utterance", "roughly one week ago to one month ago", "roughly one to two months ago up to one or two years ago" or in the "distant or legendary past".

[9] An inflected verb means a variation of the verb by means of an affix.

[10] We did not consider cases in which adjectives do not modify nouns, occurring as predicates in internally headed relative clauses. Only 4/1366 languages: Seri (Mexico, 518 speakers); Tiipay (jamul, Kumiai) (Mexico, USA, 330 speakers); Kutenai (Canada, USA, 12 speakers); Choctow (USA, 11390 speakers) exists in our sample.

rather than morphosyntactic, terms. A reason construction is regarded as one encoding a causal relation between two events, such that one of the two (the event coded by the reason clause, or the dependent event) represents the reason for the other event (the main event) to take place.." (from WALS). In particular, we build a categorical variable equal to 0, 0,5 and 1 depending on whether the language has (a) Balanced or (b) both Balanced and deranked, or (c) Deranked respectively verb form in reason clauses.[11]

7. Exceed Comparative Constructions As marker of expressions of comparisons we use information on quantitative comparative constructions, that allow to rank and compare things. These "essentially involves three things: a predicative scale, which, in language, is usually encoded as a gradable predicate, and two objects". In particular, we use a dummy variable for whether a language has exceed comparative particles (1) or not (0).[12]

### 2.2.2 Weighting of grammatical structure variables

To take into account the linguistic diversity within countries, we construct weighted measures of our grammatical structure variables described above. When there is no missing data, our measure is a simple average where the weights correspond to the percentage of speakers for which the given language is their native language. As detailed later in this section, our data coverage is high. Yet, there are cases where we do not have information on the language spoken by all the people of the country. When this is so we condition our weights on the share of the population for which we have information. footnoteWhen the data coverage is too low, computing a weighted variable might seem insignificant. Our data set allows the researcher to choose an appropriate threshold of data availability and conduct robustness checks with different thresholds. A detailed example is included in the appendix.

[11] "A balanced verb form is one that can occur in an independent declarative clause"(WALS chapter 127). That is, outside of the reason construction, as in "I couldn't go to Paris last week [because all the trains were booked out]." A deranked verb form is one that cannot be used in independent declarative clauses, as in "[Being so busy], I couldn't do anything else." That is, "a deranked verb form may lack some or all of the categorial distinctions relevant to verbs in the language (such as tense, aspect, mood or person agreement distinctions), or display special markers not used in independent clauses, e.g. special tense, aspect, mood or person markers, nominalizers, case markers or adpositions."

[12] "Exceed Comparatives have as their characteristic that the standard NP (noun phrase) is constructed as the direct object of a transitive verb with the meaning 'to exceed' or 'to surpass'. "

| Table 1: Descriptive Statistics of weighted (by number of speakers) grammatical structure variables | | | | | |
|---|---|---|---|---|---|
| | N | Minimum | Maximum | Mean | Std. Deviation |
| Number of Genders | 204 | 0 | 5 | 1.93 | 1.50 |
| The Past Tense | 203 | 0 | 2 | .84 | .44 |
| The Optative | 203 | 0 | 1 | .08 | .23 |
| Order of Object and Verb | 204 | 0 | 1 | .79 | .35 |
| Order of Adjective and Noun | 204 | 0 | 1 | .56 | .45 |
| Reason Clauses | 204 | 0 | 1 | .39 | .29 |
| E.Comparative | 202 | 0 | 1 | .14 | .32 |

| Table 2: Correlations of weighted grammatical structure variables | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Number of Genders | The Past Tense | The Optative | Order of Object and Verb | Order of Adjective and Noun | Reason Clauses | E.Comparative |
| Number of Genders | 1 | | | | | | |
| The Past Tense | 0.47*** | 1 | | | | | |
| The Optative | -0.13** | -.021 | 1 | | | | |
| Order of Object and Verb | .093 | -0.11* | -0.32*** | 1 | | | |
| Order of Adjective and Noun | 0.18*** | .085 | -0.15** | 0.25*** | 1 | | |
| Reason Clauses | 0.42*** | 0.29*** | -.040 | -.043 | .104 | 1 | |
| E.Comparative | 0.28*** | -0.12* | -0.13* | .107 | 0.23*** | .102 | 1 |

## 2.3 Descriptive Statistics

Table 1 presents the descriptive statistics of our seven weighted (by number of speakers) grammatical structure variables for our sample of 204 countries.[13] Table 2 presents the correlation of our weighted grammatical structure variables. A bit less than half (8/21) are not significantly correlated, and overall, correlations across grammatical variables are relatively small. This means that they can be used simultaneously in regression analysis without suffering severe colinearity problems.

### 2.3.1 Data coverage

In terms of the percentage of the population and the languages they speak, our data coverage is very high. On average we cover 92% of the population per country. The remaining 8% results largely from the fact that we only consider languages whose speakers amounted for more than 1% of the country's population and also from the fact that a few countries have limited data availability.[14][15] In terms of categorizing the grammatical structures of all the languages spoken in the countries of the world, 50% of the data is available. This is due to the fact that our source (WALS) is an ongoing effort from linguistic researchers and it is still incomplete. This is not a concern since the languages for which data is not available are the ones spoken by very few people and we are only taking into account languages spoken by 1% of the country's population.

## 2.4 Economic, Gender and Institutions data set

We use three data sources for our economic, gender and institutional variables:

1. World Development Indicators (WDI) database. The WDI is an aggregate database that has data from many databases.[16]

[13] For three of our seven variables one or two countries lacked information on these grammatical features.

[14] We couldn't account for the whole population in highly fragmented countries, such as Papua New Guinea, where there is more than 800 living languages with a mean number of speakers per language of 1,200.

[15] We were able to find data for only 47% of the population of the Democratic Republic of Congo.

[16] This allows us to use different sources for measuring the same variable as a robustness check.

2. PRC database. It includes data from two distinct methodologies: Political Risk Services and International Country Risk Guide.[17].

3. Institutional Profiles Database (IPD) 2009. It presents indicators on institutional characteristics of 123 developed and developing countries.[18]

All of our variables are based on a country average based on annual data from 1990-2010 period. Our economic variables capture income distribution variables (GINI, and the income share held by highest 10%) and gross savings as % of GDP. Recent research has shown that cultural values are also important in determining the inclination to save (Guiso, Sapienza Zingales, 2006). For example, cultural differences (the extent thriftiness taught to children), has been shown to be important in explaining cross-country variation in saving behavior (Guiso et al., 2006). Our gender variables include female educational, household, firm ownership and labor market related variables. Based on the literature, cultural values regarding gender influence female economic decisions and prominently among them labor force participation (Hofstede 1980, 1983; House at al 2004). Regarding our institutional variables, one of the most salient institutions in the literature is property rights. North and Weingast (1989) argue that sound property rights and incentive schemes made possible Britains distinctive institutions that enabled it to industrialize first. North, Summerhill Weingast (2000) ascribe economic failures in Latin America to a system of shared beliefs about political legitimacy and individual rights, which has blocked the institutions that might guarantee property rights and contract enforcement. We included property rights variables such as the Strength of legal rights index, among other variables capturing formal institutions (rules and regulations) (La Porta et al.1997, 1998). The economics literature has also pointed out to the importance of informal institutions and the influence of culture on it. (Boettke et al, 2008; Leeson, 2005; Xu and Shenkar (2002) Gaur and Lu (2007) Williamson, 2009). Informal

[17]Please refer to http://www.prsgroup.com/

[18]These refer to 1. Political institutions; 2. Safety, Law and Order, Control of violence; 3. Functioning of Public administrations; 4. Free Operation of Markets; 5. Coordination of actors, Strategic vision, Innovation; 6. Security of transactions and contracts; 7. Market regulations, Social dialogue; 8. Openness to the outside world; 9. Social cohesion and mobility. It was built by researchers from the French Ministry for the Economy, Industry and Employment (MINEIE) and the French Development Agency (AFD) and was constructed from a world survey conducted with MINEIE and AFD agencies present in the countries covered in the database. Please refer to http://www.cepii.fr/ProfilsInstitutionnelsDatabase.htm

institutions include norms and codes of behavior (Boettke et al, 2008; Leeson, 2005; Xu and Shenkar (2002) Gaur and Lu (2007) Williamson, 2009). To capture these, we include measures capturing corruption practices, such as bribery and tax evasion (whether firms are expected to give gifts in meetings with tax officials and whether they fail to report all sales for tax purposes).

### 2.5 Sample

We collected grammatical structure and economic, gender and institutions data for a sample of 204 countries, which constitutes the largest data set on markers of culture available.

## 3 Correlation between Grammar and Socio-Economic Structures

The following tables present the correlation between grammatical structures and economic, gender and institutional variables.[19] The correlation matrixs show a wide range of significant correlations between the grammatical structure variables and economic, gender and institutional variables. Some culture dimensions are more relevant to a specific socioeconomic or institutional variable (Shenkar 2001) and therefore, not all the cultural variables should relate to all real activity variables. The correlation matrixes present some very interesting results which we discuss next. Because these correlations have never been studied before, there is no a priori knowledge to interpret them. If a common cultural interpretation of a grammatical structure variable emerges across correlations, we cannot reject that the feature is a marker of such cultural orientation. We summarize our findings at the end of this section, useful to guide future empirical research using data on grammatical structure variables as markers of culture, which are clear to define, measure and exogenous to current socio-economic outcomes.

[19]The number in the table is the correlation outcome. We use denote $P-Value < 0.1$, ** to denote $P-Value < 0.05$ and *** to denote $P-Value < 0.01$

| Table 3: Grammatical and Economic Structures | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Number of Genders | The Past Tense | The Optative | Order of Object and Verb | Order of Adjective and Noun | Reason Clauses | E.Comparative |
| GINI index | 0.32*** | 0.21** | -.077 | 0.16* | 0.52*** | 0.15* | 0.21** |
| Gross savings (% of GDP) | -.117 | -.062 | -.011 | .072 | -0.15* | -.041 | -.093 |
| Income share held by highest 10% | 0.27*** | 0.19** | -.065 | .137 | 0.53*** | .116 | 0.22*** |

### 3.1 Grammatical and Economic Structures

In general, **inequality** measures (the Gini index and the income share held by the richest 10% of the population) are very significantly and positively correlated with number of genders, the past tence, order of adjective and noun and the exceed comparative variables. That is, countries where the population speaks languages that have more markers of gender, more markers of past and remoteness integrated in the grammar, that have exceed comparative particles to rank objects, and that put the noun before the adjective exhibit more unequal income distribution. According to these correlations, gender marking ( which involves different agreements dependent on nouns of different types, generally sex based), can be interpreted as a marker of societies that stress differences based on female-male and other distinctions such as social status or age and are therefore more rigid. Languages with higher number of genders reflect, then, a more hierarchical (less egalitarian) culture. Regarding languages with past tense marking (and with higher degrees of remoteness available) these may reflect cultures where tradition and inheritance is more important. Further, languages where exceed comparative particles (that allow to rank and compare things) are present may reflect cultures that are more competition oriented. Finally, languages with noun-adjective order may reflect societies that are less flexible or less innovation oriented. On the other hand, it is striking how our measure of **savings** is not strongly significantly correlated with any of our grammatical structure variables.

### 3.2 Grammatical and Gender Structures

We analyze the relation between grammatical and gender related outcomes due to the fact that an existing body of research (Hofstede 1980, 1983 and House et al, 2004) has studied how gender dimensions are an important part of culture.[20] Further, gender differences can be impacted by non gender related cultural values and vice versa (Brock et al, 2008). Table 4 shows how the variables related to female literacy rates, school enrollment and relative enrollment (with respect

[20]For an elaboration on gender and culture refer to Emrich, Denmark, and Den Hartog, 2004.House and Javidan, 2004: 12 define gender egalitarianism as the degree to which an organization or a society minimizes gender role differences while promoting gender equality.

to male) are strongly negatively correlated to number of genders. That is, languages that mark more distinctions of gender invest less in **female population human capital**. Interestingly, despite the high correlation between number of genders and the past tense, note how the past tense is not significantly correlated with any but one gender outcome variable. In particular, countries where the languages spoken have more markers of the past (the past tense variable) have a higher proportion of **female headed households**.

| Table 4: Grammatical Structure and Gender | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Number of Genders | The Past Tense | The Optative | Order of Object and Verb | Order of Adjective and Noun | Reason Clauses | E. Comparative |
| Employers, female (% of employment) | .149 | .096 | -.043 | 0.18** | .021 | .138 | 0.25* |
| Employment to population ratio, 15+, female (%) | .098 | -.087 | -.113 | .039 | .019 | -.099 | 0.47*** |
| Female headed households (% of households with a female head) | .132 | 0.26** | -0.20* | 0.26** | -.041 | .139 | 0.12 |
| Firms with female participation in ownership (% of firms) | -.125 | .021 | -.084 | 0.30*** | 0.19** | -.144 | -0.28* |
| Labor force with primary education, female (% of female labor force) | .012 | .136 | -.064 | -.013 | 0.28*** | -.087 | -.180 |
| Labor force with secondary education, female (% of female labor force) | -0.20** | -.102 | .030 | .080 | -0.47*** | -.150 | -.040 |
| Labor force with tertiary education, female (% of female labor force) | .098 | .126 | -.056 | 0.18* | -0.28*** | .012 | -.160 |
| Labor participation rate, female (% of female population ages 15+) | .120 | -.055 | -.127 | .056 | -.035 | -.075 | 0.41*** |
| Literacy rate, adult female (% of females ages 15 and above) | -0.20** | .070 | .026 | 0.24*** | -0.28*** | -0.14* | -0.32** |
| Literacy rate, youth female (% of females ages 15-24) | -0.24*** | .035 | .027 | 0.24*** | -0.22*** | -0.18** | -0.28** |
| Net intake rate in grade 1, female (% of official school-age population) | -0.32*** | -.069 | .013 | .108 | -0.35*** | -.126 | -0.33** |
| Part time employment, female (% of total female employment) | .163 | .108 | 0.33*** | -0.40*** | -.005 | .118 | -.190 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Persistence to grade 5, female (% of cohort) | -.045 | .086 | -0.14* | 0.17** | -0.25*** | -.030 | -0.30** |
| Persistence to last grade of primary, female (% of cohort) | -0.16** | .043 | .012 | .048 | -0.37*** | -.055 | -0.38*** |
| Ratio of female to male primary enrollment (%) | -0.14* | .105 | -0.13* | 0.35*** | -0.20*** | -.121 | -0.32*** |
| Ratio of female to male secondary enrollment (%) | -0.15** | .119 | -.087 | 0.29*** | -0.21*** | -.069 | -0.41*** |
| Ratio of female to male tertiary enrollment (%) | -0.16** | .038 | .051 | 0.22*** | -0.23*** | -.057 | -0.29** |
| Ratio of young literate females to males (% ages 15-24) | -0.15* | .106 | -.019 | 0.27*** | -0.15* | -.117 | -0.22* |
| Repeaters, primary, female (% of female enrollment) | .071 | .000 | -.092 | -.012 | 0.43*** | .021 | .110 |
| Repeaters, secondary, female (% of female enrollment) | .119 | .055 | -.119 | -.026 | 0.38*** | .108 | .070 |
| School enrollment, primary, female (% net) | -0.23*** | -.019 | -.025 | 0.32*** | -0.29*** | -0.16** | -0.29** |
| School enrollment, secondary, female (% net) | -0.19** | .030 | -.042 | 0.21*** | -0.45*** | -.080 | -0.49*** |
| School enrollment, tertiary, female (% net) | -0.18** | -.040 | -.004 | 0.19** | -0.42*** | -.073 | -0.39*** |
| Secondary education, general pupils (% female) | -0.18** | .117 | -.059 | 0.26*** | -0.24*** | -.098 | -0.36*** |
| Secondary education, teachers (% female) | -0.14* | .039 | .082 | 0.25*** | -0.32*** | -.057 | -0.32*** |
| Self-employed, female (% of females employed) | 0.23*** | -.038 | .133 | -.055 | 0.32*** | -.073 | 0.42*** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Unemployment, female (% of female labor force) | .065 | .140 | 0.22*** | .037 | .023 | .057 | -0.15 |
| Vulnerable employment, female (% of female employment) | .071 | -.066 | 0.25*** | -0.25*** | 0.33*** | -.077 | 0.38*** |

### 3.3 Grammatical and Institutional Structures

As the table shows, political, legal, financial institutions are strongly correlated with grammatical structures. Given the size of correlation, we do not attempt to discuss each of them. Regarding the **quality of political institutions**, they are worst in countries where the population speaks languages with the noun preceding the adjective, and better in those where they speak languages where the verb precedes the object. The next section summarizes the correlations and interpret them.

Table 5: Grammatical Structure and Institutions

| | Number of Genders | The Past Tense | The Optative | Order of Object and Verb | Order of Adjective and Noun | Reason Clauses | E.Comparative |
|---|---|---|---|---|---|---|---|
| Security of transactions and contracts | -.069 | .039 | -.134 | .136 | -0.45*** | .051 | -0.23** |
| Government respect for contracts | -.004 | 0.16* | -0.18** | .042 | -0.18** | .118 | -0.19** |
| Frequency of bankruptcy | .038 | .075 | -0.16* | .075 | -.017 | .023 | -.041 |
| Enforcement of bankruptcy law | -.007 | -.026 | -.049 | -.122 | .055 | -.063 | -.014 |
| Information on G&S markets | -.137 | .079 | -0.21** | 0.23*** | -0.41*** | -.059 | -0.26*** |
| Rural land tenure: traditional property | 0.19** | -.069 | .079 | -.048 | 0.33*** | .077 | .138 |
| Rural land tenure: public property | .081 | -.008 | -.131 | .091 | 0.16* | -.054 | 0.21** |
| Diversity of land tenure rights systems | -.118 | .047 | .086 | .069 | -0.40*** | -.033 | -0.23** |
| Government recognition of diversity of land tenure rights systems | -0.19** | -.064 | .119 | .062 | -0.34*** | -.001 | -0.18** |
| Land tenure: security of ownership | -0.16* | .006 | -.021 | .083 | -0.45*** | .005 | -0.28*** |
| Land tenure: demand for land | -.125 | -.057 | -.079 | -.038 | .047 | -.111 | .089 |
| Land tenure and large investors | -.090 | -.107 | -.090 | 0.15* | -.004 | -.064 | .064 |
| Respect for workers' rights | -.118 | .090 | -0.22** | 0.23** | -0.30*** | -.095 | -0.26*** |
| Weak employment contrat rigidity | .005 | .013 | .134 | -.101 | -.083 | .025 | .124 |
| Financial information | -.017 | .098 | -.121 | .136 | -0.36*** | .023 | -0.23*** |
| Privatizations in the financial sector | .025 | .095 | -.077 | .011 | -.037 | -0.17** | .078 |
| Nationalizations in the financial sector | -.054 | 0.17** | -.125 | -.054 | -0.19** | .026 | -0.15* |
| Freedom in the allocation of loans | .027 | .011 | -.090 | .134 | -0.24*** | .041 | -0.20** |
| Competence of bank executives | .010 | .058 | -.121 | 0.16* | -0.31*** | .082 | -0.21** |
| Importance of venture capital | -.128 | -.042 | -0.20** | 0.20** | -0.27*** | -.018 | -.053 |
| Sovereign wealth fund policy | -.051 | .013 | .061 | .045 | -.043 | .036 | -.021 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Competition within the banking system | -.017 | .029 | -.066 | -.038 | -0.21** | .061 | -.058 |
| Regulation of competition in banking | -.097 | 0.19** | -0.17** | .091 | -0.42*** | -.023 | -0.24*** |
| Monitoring and auditing in banking | -.033 | -.023 | -.062 | 0.25*** | -0.35*** | .012 | -0.23** |
| Reform of financial regulations | .005 | .042 | -.133 | .032 | -.032 | -.064 | .058 |
| Financial openness | -.097 | -.024 | -0.17* | 0.18** | -0.24*** | -.001 | -0.22** |
| Bureaucracy Quality | -0.20** | -.066 | -0.19** | 0.17** | -0.46*** | -.133 | -0.17* |
| Contract Viability | -0.27*** | -.036 | -.102 | 0.20** | -0.34*** | -.075 | -0.27*** |
| Democratic Accountability (K) | -0.14* | .044 | -0.19** | 0.18** | -0.43*** | -.098 | -0.31*** |
| Government Stability (A) | -0.14* | -.083 | -.055 | 0.17** | -0.20** | -.113 | .124 |
| Law & Order (I) | -0.20** | -.027 | -0.17* | 0.18** | -0.40*** | -0.18** | -0.27** |
| Legislative Strength | .062 | .026 | -.021 | -.063 | .090 | .036 | .118 |
| Burden of customs procedure, WEF (1=extremely inefficient to 7=extremely efficient) | -.015 | .048 | -0.16* | 0.16* | -0.20** | .086 | -.094 |
| Children out of school, primary | -.025 | -.064 | .045 | -0.24*** | -.094 | .039 | 0.15** |
| Cost of business start-up procedures (% of GNI per capita) | 0.30*** | -.034 | -.079 | -.078 | 0.29*** | .117 | 0.30*** |
| Ease of doing business index (1=most business-friendly regulations) | 0.14* | -.086 | .045 | -0.16** | 0.34*** | .027 | 0.20*** |
| Firing cost (weeks of wages) | 0.14* | .045 | -.053 | -.008 | 0.23*** | .134 | 0.23*** |
| Firms expected to give gifts in meetings with tax officials (% of firms) | -.131 | -.066 | .148 | -0.28*** | -0.28*** | -.143 | .030 |
| Firms that do not report all sales for tax purposes (% of firms) | 0.23** | .025 | .095 | -.058 | 0.22** | .136 | 0.33*** |
| Informal payments to public officials (% of firms) | .153 | -.047 | -.003 | -.129 | .044 | .039 | 0.34*** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Logistics performance index: Efficiency of customs clearance process (1=low to 5=high) | -.131 | -.028 | -0.23*** | 0.22*** | -0.29*** | -.013 | -.128 |
| Out-of-pocket health expenditure (% of private expenditure on health) | -0.14* | -0.13* | 0.14* | -0.14* | -0.18* | -.072 | -.091 |
| Out-of-pocket health expenditure (% of total expenditure on health) | .002 | -.062 | 0.16** | -0.27*** | .046 | .038 | 0.20** |
| Procedures to enforce a contract (number) | -.004 | -.076 | 0.25*** | -0.15** | 0.20** | .075 | .004 |
| Procedures to register property (number) | .049 | .065 | .017 | -.073 | 0.17** | -.011 | .024 |
| Start-up procedures to register a business (number) | .023 | -.106 | .076 | -.027 | 0.31*** | -.084 | 0.14** |
| Strength of legal rights index (0=weak to 10=strong) | .101 | 0.15* | -.023 | 0.21*** | -0.28*** | .026 | .095 |

### 3.4 Grammar and Culture: Summary and Interpretation

These correlations have never been studied before, there is no a priori knowledge to interpret them. We highlight the following most striking correlation patterns and out interpretation of them. The Number of genders (Genders) is negatively correlated with female education (literacy and schooling). Gender marking is stronger in societies that stress differences based on female-male and other distinctions such as status or age and are therefore more rigid. Languages with higher number of genders reflect, then, a more hierarchical (less egalitarian) culture. The past tense is positive correlated with economic inequality and female headed households. Therefore, languages with higher past tense marking may reflect value of tradition and inheritance. The optative is negatively correlated with the quality of political institutions, respect of workers rights. It is positively correlated with part time female employment and unemployment. It seems to capture the inclination to respect rules versus individual interest. The order of adjective and noun in particular, when the noun precedes the adjective is positively correlated with economic inequality and worst female outcomes. Mixed results for corruption, worst political institutional quality and business environment. These points us towards interpreting languages where the noun precedes the adjective as less open and flexible, with lower value for entrepreneurship. The order of object and verb, in particular, when the verb precedes the object, is positively correlated with female literacy and schooling and quality political institutions. Negatively with gift giving to tax officials. We interpret the order of object and verb as denoting the value of effort versus status in wealth acquisition and economic life. Reason clause is virtually unrelated to economic, gender and institutional variables, except it is negatively correlated with law and order. We do not have an interpretation for it. Finally, languages with presence of exceed comparative particles, that allow to rank and compare things are very significantly related with many socio-economic variables. In particular, exceed comparative is positively correlated with inequality, negatively with female schooling and literacy. Positively with female labor force participation and to corruption. Negatively related to quality of institutions. We interpret it as denoting value for competition, mainly from its relation with economic inequality.

# 4 Regression Analysis

[work in progress]

# 5 Conclusion

Definition, measurement and endogeneity problems challenge the study of the relation between culture and economic development, limiting what we can learn from it. Yet, much of the variation in the development of countries results from unobservables. Such unexplained variations are often attributed to variations in the design or performance of institutions and in cultural variations. To move forward in identifying the causal impact of culture on economic outcomes and on the functioning of formal institutions, this paper proposes to use languages grammatical structures as markers of culture. These are straightforward to define, measure and are exogenous to current economic outcomes. We build a data set of seven linguistic based markers of culture, for all the languages spoken in the countries of the world, related to expressions of time, gender, among others. We construct a weighted variable that takes into account the diversity of languages and percentage of speakers within a country, and whether languages are inherited from colonization or not. We present a correlation of grammatical structures with economic, gender and institutional variables and find statistically significant correlations.

# 6 References

**Banfield**, *The moral basis of a backward society*, The University Of Chicago, Free Press, 1958.

**Boettke, Coyne, Leeson, and Sautet**, (2005). "The New Comparative Political Economy". *The Review of Austrian Economics*, (3), 281-304.

**Corbett, Greville G.. 2011**. Number of Genders. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 30. Available online at http://wals.info/chapter/30 Accessed on 2011-08-31.

**Cristofaro, Sonia. 2011**. Reason Clauses. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 127. Available online at http://wals.info/chapter/127 Accessed on 2011-08-31.

**Dahl, sten Velupillai, Viveka. 2011**. The Past Tense. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 66. Available online at http://wals.info/chapter/66 Accessed on 2011-08-31.

**Dryer, Matthew S. Haspelmath, Martin** (eds.). 2011. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. Available online at http://wals.info/ Accessed on 2011-08-27.

**Dryer, Matthew S.. 2011**. Order of Object and Verb. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 83. Available online at http://wals.info/chapter/83 Accessed on 2011-08-31.

**Dryer, Matthew S.. 2011**. Order of Adjective and Noun. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 87. Available online at http://wals.info/chapter/87 Accessed on 2011-08-31.

**Falck, Oliver Heblich, Stephan Lameli, Alfred Suedekum, Jens, 2010.** Dialects, Cultural Identity, and Economic Exchange, IZA Discussion Papers 4743, Institute for the Study of Labor (IZA).

**Gaur, A. S., and Lu, J.W.**, 2007 "Ownership strategies and survival of foreign subsidiaries: Impacts of institutional distance and experience" *Journal of Management*, 33:84-110.

**Guiso, L., Sapienza, P. and L. Zingales.** 2006. Does culture affect economic outcomes? Journal of Economic Perspectives, 20(2),2:23-48

**Guiso, Sapienza and Zingales**, 'Moral and Social Constraints to Strategic Default on Mortgages', NBER Working Paper 15145, National Bureau of Economic Research, July 2009.

**Guiso, Sapienza and Zingales**, 'Civic Capital as the Missing Link,' NBER Working Papers 15845, National Bureau of Economic Research, March 2010.

**Hofstede, G** 1980 "Culture's consequences: International differences in work-related values" *Newbury Park, CA: Sage.*

, G. 1983 "The cultural relativity of organizational practices and theories" *Journal of International Business Studies* Fall: 75-89.

**La Porta, Lopez de Silanes and Shleifer**, 'Trust in Large Organizations,' *American Economic Review*, May 1997, *87* (2), 333-38.

**La Porta, Lopez-de-Silanes, Shleifer, and Vishny**. 1998. "Law and finance." *Journal of Political Economy*, 106(6): 1113-55.

**Leeson, P. T.** (2005)." Endogenizing fractionalization" *Journal of Institutional Economics*, 1: 75-98.

**Licht, Goldschmidt and Schwartz**, 'Culture, Law, and Corporate Governance', *International Review of Law and Economics*, June 2005, *25* (2), 229-255.

**Licht, Goldschmidt, and Schwartz**, 'Culture Rules: The Foundations of the Rule of Law and Other Norms of Governance,' *Journal of Comparative Economics*, December 2007, *35* (4), 659-688.

**North**, *Institutions, institutional change, and economic performance*, Cambridge University Press, 1990.

**North, D. C., Weingast, B. R.** 1989."Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England". *The Journal of Economic History*, 49(04), 803.

**North, D. C., Summerhill, W., and Weingast, B. R.** 2000. "Order, Disorder, and Economic

Change", in B. B. de Mesquita and H. Root, eds., Governing for prosperity New Haven. pp. 1758.

**Dobrushina, Nina van der Auwera, Johan Goussev, Valentin. 2011** The Optative. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 73. Available online at http://wals.info/chapter/73 Accessed on 2011-08-31.

**Putnam, Leonardi and Nanetti**, *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton paperbacks, Princeton University Press, 1993.

**Stassen, Leon. 2011**. Comparative Constructions. In: Dryer, Matthew S. Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 121. Available online at http://wals.info/chapter/121 Accessed on 2011-09-01.

**Tabellini**, 'Culture and Institutions: Economic Development in the Regions of Europe', CESifo Working Paper Series 1492, CESifo Group Munich, 2005.

**Tabellini**, 'Presidential Address Institutions and Culture,' *Journal of the European Economic Association*, May 2008, *6* (2-3), 255-294.

**Tabellini**, 'The Scope of Cooperation: Values and Incentives,' *The Quarterly Journal of Economics*, August 2008, *123* (3), 905-950.

**Tang and Koveos** 'A Framework to Update Hofstede's Cultural Value Indices: Economic Dynamics and Institutional Stability,' *Journal of International Business Studies*, 2008, *39* (6), 1045-1063.

**Williamson, C.** (2009). "Informal Institutions Rule: Institutional Arrangements and Economic Performance". *Public Choice*, (3), 371-387

**Xu, D., and Shenkar, O.**, 2002 "Institutional distance and the multinational enterprise"] *Academy of Management Review*, 27(4): 608-618.

Ashkanasy, N., Gupta. V., Mayfield, M. S. and Trevor-Roberts, E. (2004). Future Orientation. In House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., and Gupta. V. (Eds). Leadership, culture, and organizations: The GLOBE study of 62 societies. Thousand Oaks, CA: Sage Publications, pp. 396-436

Brock, D. M., Shenkar, O., Shoham, A., and Siscovick, I. C., 2008 National Culture and Expatriate Deployment. Journal of International Business Studies 39(8): 1293-1309.

Emrich, G.C., Denmark, F.L., and Den Hartog, D.H., 2004 Cross cultural Differences in Gender Egalitarianism: Implication for Societies, Organizations, and leaders In R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman and V. Gupta (Eds), Leadership, culture, and organizations: The GLOBE study of 62 societies: 343394. Thousand Oaks, CA: Sage Publications.

Hofstede, G. 1980. Cultures consequences: International differences in work-related values. Newbury Park, CA: Sage.

Hofstede, G. 1983. The cultural relativity of organizational practices and theories. Journal of International Business Studies, 14(2): 7589.

Hofstede, G. and Bond, M. H. (1988) The Confucius Connection: From Cultural Roots to Economic Growth, Organizational Dynamics 16(4): 5-21.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., Gupta, V. 2004. Leadership, culture, and organizations: The GLOBE study of 62 societies.. Thousand Oaks, CA: Sage Publications.

House, R. J., Javidan, M. 2004. Overview of GLOBE. In R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman V. Gupta (Eds), Leadership, culture, and organizations: The GLOBE study of 62 societies: 928. Thousand Oaks, CA: Sage Publications.

Minkov, M., and Hofstede, G. (2011). Hofstede's fifth dimension; New evidence from the World Values Survey. Journal of Cross-Cultural Psychology (forthcoming).

Shenkar, O. 2001. Cultural distance revisited: Towards a more rigorous conceptualization and measurement of cultural differences. Journal of International Business Studies, 32(3):519535.

# 7 Appendix

## 7.1 Data collection challenges

### 7.1.1 Which languages are relevant for African countries?

Africa has the particularity of being extremely linguistically diverse (for instance, there are about 279 living languages in Cameroon and 134 in Sudan). Hence, in many African countries, people learn and use a vehicular language as a second language to communicate with people from different ethnies. For example, there are 74 languages in Kenya, and the most widely spoken maternal language, Kikuyu, covers only 18% of the population (the diversity index for Kenya is 0.877). Nevertheless, about 40% of the population know a form of Swahili which is used as a vehicular language even though it is the maternal language of only 0.05% of the population. Hence, should we account for vehicular languages as a country's dominant language? We didn't make this choice because these vehicular languages are almost always secondary languages, not always fully known (people learn the basics to understand each other) and are usually mixed with their maternal language. We thus chose to only account for maternal languages. For the same reasons, colonial languages such as French or Portuguese have been given a small weight in the dataset: even if they are the official languages in many African countries, they are usually only used in official documents, and only the highly educated citizens are able to speak them.

### 7.1.2 The problem of Creoles in the Caribbeans

Many countries in the Caribbeans have a Creole as a dominant language. For instance, in Saint Kitts and nevis, 98% of the population speaks the Saint Kitts English Creole. However, each variety of Creole is specific to a country or a region, so that all the varieties of Creoles cannot be accounted as one single language. The main issue then is that there is no information on WALS on the majority of these Creoles, so that we cannot deal with a lot of Caribbean islands.

### 7.2 Detailed weighting when missing data

For each country we compute the availability of the data on the maternal language of a country's population, which we call "data coverage".[21] When all data is available our weights are simply what we call Language coverage variable: the share of the population of a country that speak a specific language as their maternal language.For instance, if in country C 250 000 people speak language L1, then the Language coverage (L1)=25

### 7.3 Coding of dataset

There are sev grammatical structure variables in the dataset, each one corresponding to a specific feature from the WALS database. We also include in parenthesis the stability index (very stable, stable, unstable, very unstable) derived by Wichmann Holmanlni paper.[22]

1. Number of Genders (see http://wals.info/chapter/30).(very stable)

| WALS | None | Two | Three | Four | Five or more |
|---|---|---|---|---|---|
| Dataset | 0 | 2 | 3 | 4 | 5 |

2. Chapter 66: the Past Tense (see http://wals.info/chapter/66) (very stable)

| WALS | No grammatical marking of past/non-past distinction | Past/non-past distinction marked; no remoteness distinction | Past/non-past distinction marked; 2-3 degrees of remoteness distinguished | Past/non-past distinction marked; at least 4 degrees of remoteness distinguished |
|---|---|---|---|---|
| Dataset | 0 | 1 | 2 | 3 |

[21]For instance, in a country C with 1 million people, if we find the maternal language variables of only 500 000 people, then Data coverage=50%.

[22]Their results, based on their quantitative measure to assess features stability, have been matched to categorical statements made in the literature and coincide in high percentage with these.

3. Chapter 73: the Optative (see http://wals.info/chapter/73) (very stable)

| WALS | Inflectional optative absent | Inflectional optative present |
|---|---|---|
| Dataset | 0 | 1 |

4. Chapter 83: Order of Object and Verb (see http://wals.info/chapter/83)(very stable)

| WALS | Both orders with neither order dominant | Object precedes verb (OV) | Object follows verb (VO) |
|---|---|---|---|
| Dataset | 0.5 | 0 | 1 |

5. Chapter 87: Order of Adjective and Noun (see http://wals.info/chapter/ 87) (very stable)

| WALS | Both orders of noun and modifying adjective occur, with neither dominant | Modifying adjective precedes noun (AdjN) | Modifying adjective follows noun (NAdj) |
|---|---|---|---|
| Dataset | 0.5 | 0 | 1 |

We dropped the following category: Adjectives do not modify nouns, occurring as predicates in internally headed relative clauses. Only 4/1366 languages: Seri (Mexico, 518 speakers); Tiipay (jamul, Kumiai) (Mexico, USA, 330 speakers); Kutenai (Canada, USA, 12 speakers); Choctow (USA, 11390 speakers) Source: The Ethnologue, Languages of the World (Accessed on 13/07/ 11).

6. Chapter 127: Reason Clauses (see http://wals.info/chapter/127) (stable)

| WALS | Balanced | Balanced/deranked | Deranked |
|---|---|---|---|
| Dataset | 0 | 0.5 | 1 |

7. Chapter 121: Comparative Constructions (see http://wals.info/ chapter/121) (very stable)

| | | |
|---|---|---|
| WALS Exceed Compar-ative | | |
| Dataset (is the variable present?) | If yes: 1; If not: 0. | |